Letter to the Editor.

Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable

Carla Mavian[1,2#], Sergei Kosakovsky Pond[3], Simone Marini[1,4], Brittany Rife Magalis[1,2], Anne-Mieke Vandamme[5,6], Simon Dellicour[5,7], Samuel V. Scarpino[8], Charlotte Houldcroft[9], Julian Villabona-Arenas[10], Taylor K. Paisie[1,2], Nídia S. Trovão[11], Christina Boucher [12], Yun Zhang[13], Richard H. Scheuermann [14], Olivier Gascuel[15], Tommy Tsan-Yuk Lam[16], Marc A. Suchard[17], Ana Abecasis[6], Eduan Wilkinson[18], Tulio de Oliveira[18], Ana I. Bento[19], Heiko A. Schmidt[20], Darren Martin[21], James Hadfield[22], Nuno Faria[23], Nathan D. Grubaugh[24], Richard A. Neher [25],Guy Baele[5], Philippe Lemey[5], Tanja Stadler[26], Jan Albert[27], Keith A. Crandall[28], Thomas Leitner [29], Alexandros Stamatakis [30], Mattia Prosperi[1,4#], Marco Salemi[1,2#]

1. Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
2. Department of Pathology, Immunology and Laboratory medicine, University of Florida, Gainesville, FL, USA
3. Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA;
4. Department of Epidemiology, University of Florida, Gainesville, FL, USA
5. KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Clinical and Epidemiological Virology, 3000 Leuven, Belgium
6. Center for Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, 1349-008 Lisbon, Portugal
7. Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP264/3, 50 Av. FD Roosevelt, 1050 Bruxelles, Belgium
8. Network Science Institute, Northeastern University
9. Department of Medicine, University of Cambridge, Cambridge CB2 3QG, UK
10. Centre for the Mathematical Modelling of Infectious Diseases and Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK
11. Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health,16 Center Drive, Bethesda, Maryland, USA.
12. Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL, USA
13. Craig Venter Institute, La Jolla, CA 92037, USA
14. Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 92037, USA; J. Craig Venter Institute, La Jolla, CA 92037, USA; Department of Pathology, University of California, San Diego, San Diego, CA 92093, USA
15. Unité de Bioinformatique Evolutive, DBC/C3BI USR 3756 CNRS & Institut Pasteur, Paris, France
16. State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China
17. David Geffen School of Medicine at UCLA,Departments of Biomathematics, Biostatistics and Human Genetics; South Los Angeles, CA
18. KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), UKZN, Durban, South Africa
19. School of Public Health, Department of Epidemiology and Biostatistics, Indiana University, IN, USA
20. Center for Integrative Bioinformatics Vienna (CIBIV), Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna, Austria
21. Department of integrative biomedical Sciences, IIDMM, University of Cape Town, South Africa
22. Department of Medicine, University of Washington
23. Department of Zoology. University of Oxford. Oxford, UK
24. Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT
25. Biozentrum, University of Basel, Klingelbergstrasse 70, 4056 Basel, Switzerland
26. Department of Biosystems Science and Engineering of the Swiss Federal Institute of Technology (ETH Zürich) in Basel

27. Department of Microbiology Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden.
28. Department of Biostatistics & Bioinformatics, Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Washington, DC, USA
29. Theoretical Biology & Biophysics,MS K710,Los Alamos National Laboratory, Los Alamos, NM, USA
30. Heidelberg Institute for Theoretical Studies and Karlsruhe Institute of Technology, Karlsruhe, Germany

#= correspondence should be address to cmavian@ufl.edu; m.prosperi@ufl.edu; salemi@pathology.ufl.edu

Dear Editor,

There is obvious interest in gaining insights into the epidemiology and evolution of the virus that has recently emerged in humans as the cause of the coronavirus disease 2019 (COVID-19) pandemic. The recent paper by Forster *et al.* (1), analyzed 160 SARS-CoV-2 full genomes available (https://www.gisaid.org/) in early March 2020. The central claim is the identification of three main SARS-CoV-2 types, named A, B, and C, circulating in different proportions among Europeans and Americans (types A and C) and East Asian (type B). According to a median-joining network analysis, variant A is proposed to be the ancestral type because it links to the sequence of a coronavirus from bats, used as an outgroup to trace the ancestral origin of the human strains. The authors further suggest that the "ancestral Wuhan B-type virus is immunologically or environmentally adapted to a large section of the East Asian population, and may need to mutate to overcome resistance outside East Asia". There are several serious flaws with their findings and interpretation. First, and most obviously, the sequence identity between SARS-CoV-2 and the bat virus is only 96.2%, implying that these viral genomes (which are nearly 30,000 nucleotides long) differ by more than 1,000 mutations. Such a distant outgroup is unlikely to provide a reliable root for the network. Yet, strangely, the branch to the bat virus, in Figure 1 of the paper, is only 16 or 17 mutations in length. Indeed, the network seems to be mis-rooted because (see Supplementary Figure 4) a virus from Wuhan from week 0 (24th December 2019) is portrayed as a descendant of a clade of viruses collected in weeks 1-9 (presumably from many places outside China), which makes no evolutionary (2), nor epidemiological sense (3).

As for the finding of three main SARS-CoV-2 types, we must underline that finding different lineages in different countries and regions is expected with any RNA virus experiencing founder effects (2). According to Forster *et al.*'s own analysis, a single synonymous mutation (nucleotide change in a gene that does not result in a modified protein) distinguishes type A from B, while one nonsynonymous mutation (resulting in a protein with a single amino acid change) separates types A and C, and another one types B and C. Given SARS-CoV-2's fast evolutionary rate, random emergence of new mutations is entirely expected, even in a relatively short timeframe (4). When a viral strain is introduced and spreads in a new population, such random mutations can be propagated without them being selected or advantageous due to founder effects. The fact that SARS-CoV-2 sequences show some geographical clustering is not new and is nicely and interactively shown on *Nextstrain* (5), but this cannot be used as a proof of biological differences unless backed by solid experimental data (6). This is particularly true for the work of Forster *et al.* since their findings are based on a non-representative dataset of 160 genomes, with no significant correlation between prevalence of confirmed cases and number of sequenced strains per country (7, 8). The essential role of representative sampling is well documented in the literature (9), but was not acknowledged by the authors, who instead claim that their "network faithfully traces routes of infections for documented [COVID-19] cases", without taking in consideration missing viral diversity, or evaluating multiple transmission hypotheses that would be consistent with sequence data, or even providing any support on the robustness of the branching pattern in their network. Ultimately, no firm conclusion should be drawn without evaluating the probability of alternative dissemination routes.

The inappropriate application and interpretation of phylogenetic methods to analyze limited and unevenly sampled datasets begs for restraint about origin, directionality, and early clade/lineage inference of SARS-CoV-2. We feel the urgency to reframe the current debate in more rigorous scientific terms given the dangerous implications of misunderstanding the true dispersal dynamics of SARS-CoV-2 and the COVID-19 pandemic.

Acknowledgements

We are grateful to Paul Sharp, Andrew Rambaut and Nicola De Maio for their insightful suggestions and critical reading of our manuscript. CM, BRM, MP, and MS were in part supported by NSF DEB 2028221 and NIH NIAID 1R21AI138815-01.

References

1. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences 10.1073/pnas.2004999117, 202004999 (2020).
2. M. Salemi, A.-M.Vandamme, & P.Lemey, The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing. (Cambridge University Press, Cambridge, UK, 2009).
3. D. S. Hui, E. I. Azhar, T. A. Madani, F. Ntoumi, R. Kock, G. Ippolito, T. D. Mchugh, Z. A. Memish, C. Drosten, A. Zumla, E. Petersen. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China". International Journal of Infectious Diseases 91, 264–66 (2020).
4. O. G. Pybus, A. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10, 540–550 (2009).
5. J. Hadfield, C. Megill, S. M Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121-4123 (2018).
6. N.D. Grubaugh, M.E. Petrone, E.C. Holmes, We shouldn't worry when a virus mutates during disease outbreaks. *Nat Microbiol* **5,** 529–530 (2020).
7. C. Mavian, S. Marini, M. Prosperi, M. Salemi, A snapshot of SARS-CoV-2 genome availability up to 30th March, 2020 and its implications. bioRxiv 10.1101/2020.04.01.020594, 2020.2004.2001.020594 (2020).
8. S. Weaver. State of GISAID COVID-19 Sequence Availability (2020) https://observablehq.com/@stevenweaver/case-vs-sequence-count.

9. S. D. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, T. Bedford, Eight challenges in phylodynamic inference. Epidemics 10, 88-92 (2015).